Think like a spider

Main message

- Checkout Google Webmaster tools:
  - www.google.com/webmasters/tools
- Keyword search:
  - google webmaster tools

Spidering

- **Problem:** spider does not index on some dynamically generated content.
  - (See guidelines below for more examples)
- Spider has rules to index dynamic pages (ie, any URL with the "?" character).
  - It helps to keep the parameters short and few in number.
  - Parameter names like "id" or keys that look like a session id may cause the spider to skip the page
- Keep the links on a given page to a reasonable number (fewer than 100).

Solution: google *Site Index*

- Submit an XML description of URLs on your site that you want indexed

- Part of google Webmaster tools, freely available

- Webmaster tools list
  - unreachable urls
  - other errors encountered while spidering

- Limitations of site index:
  - Maximum of 50,000 URLs per site index XML file...
  - submit multiple XML files to get around this limit

Site Design Recommendations

- Use a clear hierarchy and text links.
  - Make pages reachable from static text link.
- Use a site map for users
  - links that point to the important parts of your site.
  - Break up the site map If the site map if >100 links
- Don’t use images to link to content
- Ensure the TITLE and ALT tags are descriptive
- Broken links break the spider

Site Design Recommendations

- Use of the robots.txt file on your web server.
- Dynamic pages (i.e., the URL contains a "?" character),
  - keep the number of parameters short
  - Keep parameter length short
- Keep the links on a given page to a reasonable number (fewer than 100).
- Allow search bots to crawl your sites without session IDs or arguments that track their path through the site.

  These techniques are useful for tracking users…but break bots.
- Make sure your web server supports the If-Modified-Since HTTP header.
  - Tells Google-bot if content has changed
  - Save bandwidth and overhead